2017年6月16日　統計数理研究所　オープンハウス

A Bayesian inference model for predicting transcription elongation rates with total RNA sequencing

河村　優美　　総合研究大学院大学 複合科学研究科 統計科学専攻 博士後期課程3年　（指導教員：吉田亮准教授）

1. 概要

Total RNA-sequencing without poly(A) selection（以下，Total RNA-seq）という手法を用いることで，細胞中のRNA分子（新生RNAを含む）の量を網羅的に計測することができる．本研究では，Total RNA-seqの解析から，RNA ポリメラーゼ II（以下，Pol II）の転写伸長プロセスを再構成できることを示す．Pol IIの存在確率とリードの分布の関係を状態空間モデルで表現し，ベイズ推定によりPol IIの存在確率（転写伸長の相対速度）とスプライス部位を同定することを試みる．推定された速度分布に基づき，転写伸長速度とヒストン修飾，クロマチンの状態（エピジェネティクス）を比較したところ，転写伸長速度が推定できていることが確認できた．

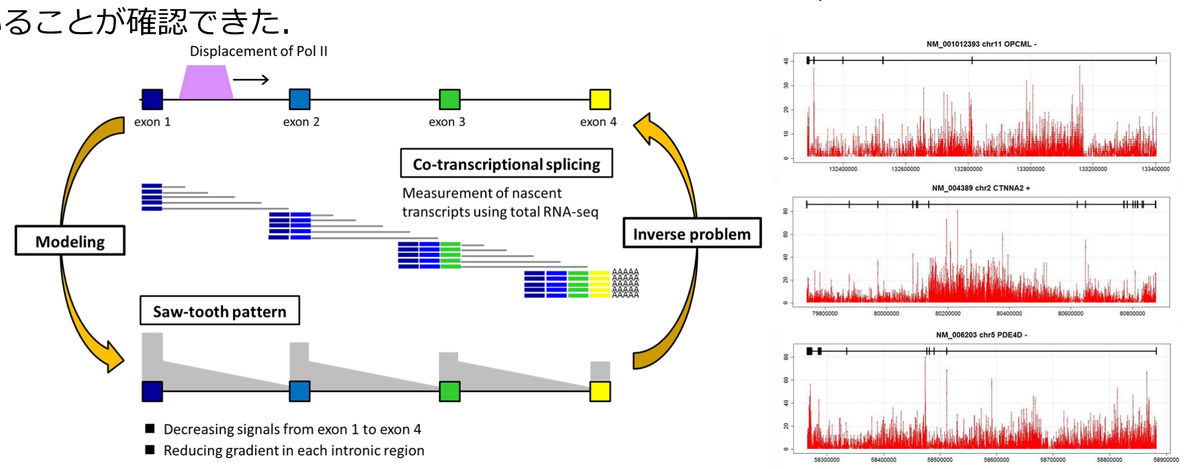


Fig1. Saw-tooth pattern from total RNA-seq

2. イントロダクション

Total RNA-seqのリード分布

Total RNA-seqのデータは，細胞中に存在するRNA分子の総量を計測したものである．データには，伸長途中のRNAも含まれる．データは，リードカウントという形式で表現される．これは，RNAの核酸塩基の位置を横軸に，個数（リード数）を縦軸にとったものである(Fig.1 右図)．Total RNA-seqの場合，リードの分布には，鋸歯状の分布が現れることが知られている．イントロン領域には，5′から3′方向に減少勾配が生じる．エキソン領域では，リード数がイントロンに比べて大きくなるため，スパイクが出現する．さらに，エキソン領域のリード数も一般的には5′から3′方向にかけて減少する．このPol IIの勾配が転写伸長速度を反映しており，このパターンをモデル化して，逆問題を解けば，転写伸長のプロセスを再構成できると考えられる(Fig.1 左図)．

これまでの転写伸長速度を測定（予測）する方法

Genome-wide run-on sequencing (GRO-seq)やnative elongating transcript sequencing (NET-seq)を使った方法がある．これにより，ゲノム上の新生転写物の位置を知ることができる．新生転写物の同定に，hidden Markov models (HMMs) のアルゴリズムを使う．これらは，転写伸長阻害剤を使って新生転写物の時系列データを複数とり，比較を繰り返すものである．一方我々の方法は，Total RNA-seqを1回行うことにより，コード領域およびnon-coding RNAにおける新生転写物を検出し，ベイズ統計を使った方法によって網羅的に転写伸長速度の計測（予測）を可能とするものである．

3. Total RNA-seqからPol II densityの推定 -1

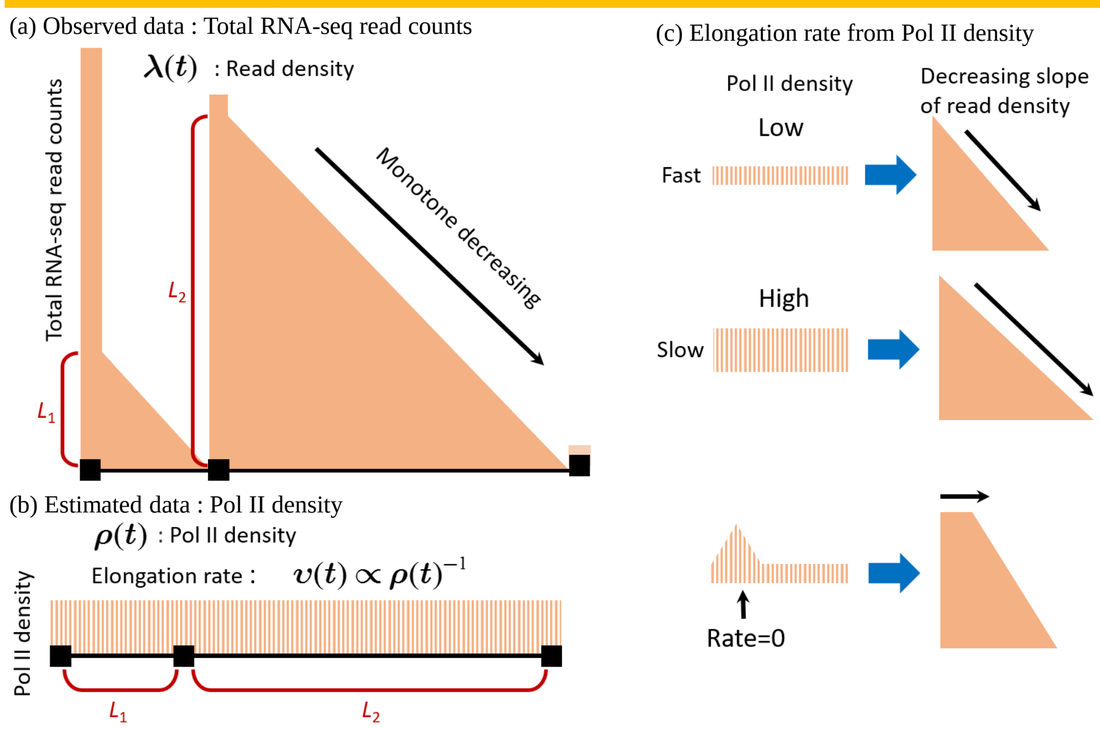
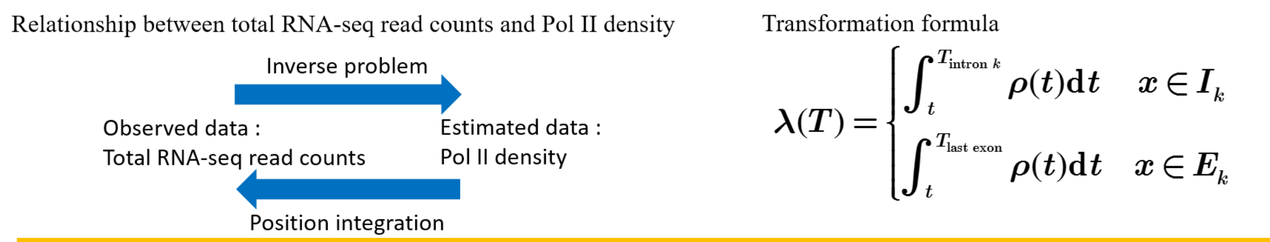


Fig2. Schematic for Pol II density from Total RNA-seq

4. Total RNA-seqからPol II densityの推定 -2

観測データであるリード分布(Fig.2 (a))から，Pol IIの存在確率を推定する(Fig.2 (b))．Fig.2で，Pol IIの存在確率とリード分布の関係を示す．横軸の遺伝子の位置 t におけるPol IIの存在確率を $\rho(t)$ とする．Pol IIの移動速度（転写伸長速度）は，存在確率の逆数に比例する(Fig.2 (b))．直観的には，リード分布の勾配が急なイントロン領域では伸長が速く（Pol IIの存在確率が小さい），勾配が緩やかな領域では速度が遅いと解釈される．これらのデータのスナップショットをリード分布にすると，伸長速度が速いところは勾配差の大きい急な減少勾配となり，遅いところは勾配差の小さい緩やかな減少勾配となる．速度が0のとき，勾配のないフラットなリード分布となる(Fig.2 (c))．また，各位置の期待リード数は$\rho(t)$の積分値で与えられる．エキソン，イントロンにおける各位置の期待リード数$\lambda(T)$は下記に示す式となる．イントロンはひとつのイントロンが終了するところでスプライスアウト（除去）するので，期待リード数$\lambda(T)$は，位置tからひとつのイントロンの終結地点までのPol IIの存在確率 $\rho(t)$ の積分値となる．エキソンはすべてのエキソンが連結し，遺伝子の終結地点でスプライスアウトされるので，期待リード数 $\lambda(T)$は，位置t から遺伝子の終結地点までのPol IIの存在確率 $\rho(t)$ の積分値となる．



5. ベイズ推定によるモデル化

我々は，ベイズ推定によるモデルを使って推定を行った(1)．Total RNA-seqのリード分布から転写伸長速度を推定する．Pol IIの存在確率とリード分布の関係を状態空間表現し，ベイズ推定によりPol IIの存在確率（転写伸長の相対速度）とスプライス部位を同時推定する方法を提案する．位置（RNAの各核酸塩基)n のリード数とPol IIの存在確率をy_n, x_nと表す．3′から5′方向に位置を表す添え字$n=1,\cdots,N$を割り当て，以下のようにy_n, x_nの状態空間表現を行う．

$$(1)\ P(x_{1:n} \mid y_{1:n}) \propto \prod_{i=1}^{n} P(y_i \mid x_{1:n}) P(x_i \mid x_{i+1})$$

$$(2)\ y_n = \lambda_n(x_{1:n}, s_n) w_n,\ \ w_n \sim \text{lognormal}(\mu, \sigma),\ \ \lambda_n(x_{1:n}, s_n) = \sum_{i=s_n}^{n} x_i \ \ (s_n \le n)$$

$$(3)\ \log x_n = \log x_{n-1} + v_n,\ \ v_n \sim \mathrm{N}(0, \gamma).$$

式(2)は観測モデルで，期待リード数 $\lambda_n(x_{1:n}, s_n)$は，s_nからnまでの状態変数の総和によって表現される．s_nは位置 n の塩基が除去される位置を表す．位置nがエキソンの場合，途中でエキソンの除去がなければ，s_nは3′末端の位置（$s_n = N$）になる．n がイントロンの場合，RSがないと仮定すれば，s_nはイントロンの終末点となる．s_nは未知パラメータであり，データから推定される．式(3)のシステムモデルは，存在確率の平滑化事前分布をあたえる．今回は，粒子フィルタを適用して，状態 x_n とスプライス部位 s_n の同時推定を試みる．ここで行った粒子フィルタは，N(3′末端の位置)から1(5′末端の位置)への逆向きの粒子フィルタリングである．

6. Pol II densityとヒストン修飾レベルの相関

ヒストンが化学修飾されることによって，遺伝子の転写活性化，不活性化などの機能が発揮される．ヒストン修飾の制御とPol IIによる転写伸長は協調していることが知られている．マウスES細胞を使って推定したPol IIと転写活性の機能を持つヒストン修飾は正の相関を示し，推定したPol IIと転写不活性の機能を持つヒストン修飾は負の相関を示す傾向にあった(Fig.3)．

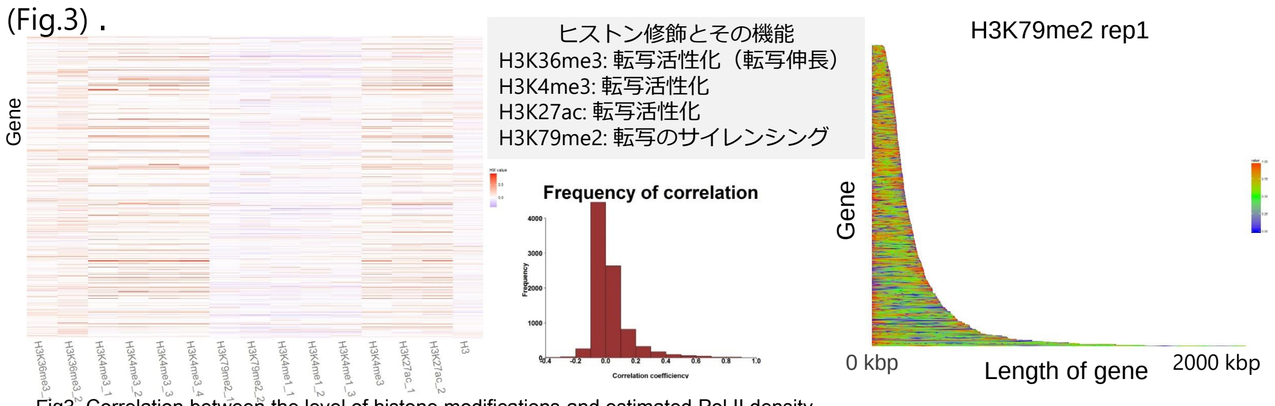


Fig3. Correlation between the level of histone modifications and estimated Pol II density

7. Pol II density（転写伸長の相対速度）とスプライス部位の推定結果

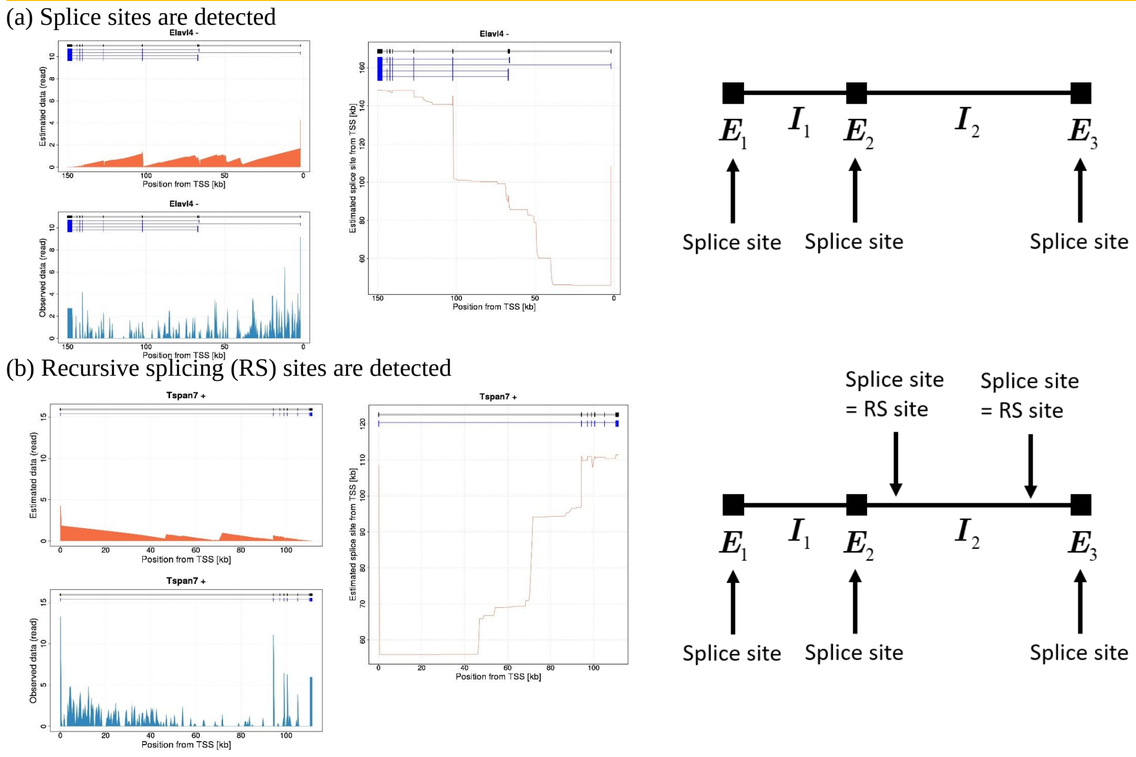


Fig4. Splice site is unknown. Our model predict unknown splice sites (recursive splicing sites).

8. Pol II densityとヌクレオソームの相関

ヌクレオソーム占有率が高いと，転写が不活性化したクロマチン状態であり，ヌクレオソーム占有率が低いと，転写が活性化したクロマチン状態である．Pol II densityが高いところは転写活性が高いことを示し、これはヌクレオソームが低いところに相当する．推定したPol IIとヌクレオソームレベルは負の相関を示す傾向にあった(Fig.5)．

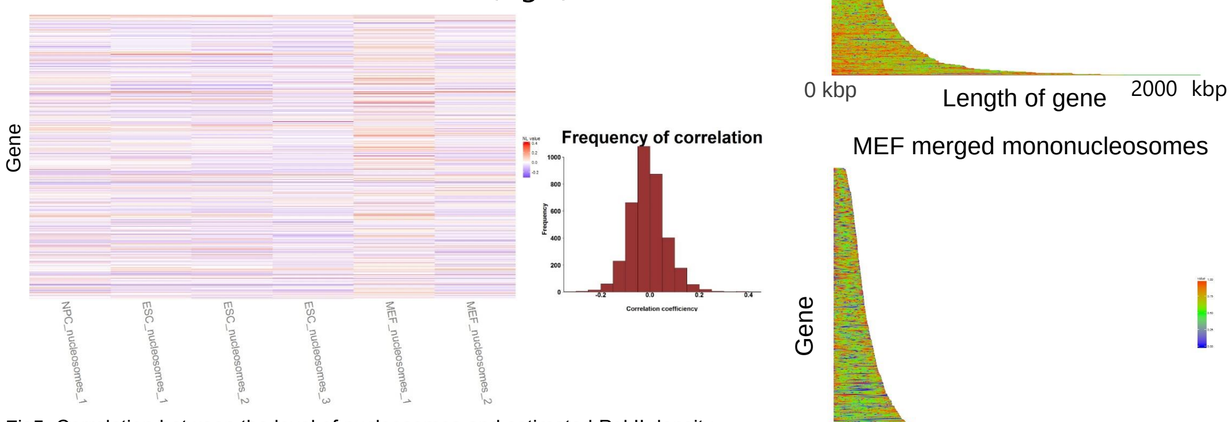


Fig5. Correlation between the level of nucleosomes and estimated Pol II density

参考文献

- Adam Ameur *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature Structural & Molecular Biology* 18, 1435–1440 (2011)
- Iris Jonkers *et al.* Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife Sciences* 3 e02407 (2014)